AD HOC WORKING GROUP ON STRATEGIC APPROACHES AND OPPORTUNITIES IN POPULATION SCIENCE, EPIDEMIOLOGY, AND DISPARITIES

REPORT ON NATIONAL CANCER INSTITUTE (NCI) EXTRAMURAL CANCER EPIDEMIOLOGY COHORT STUDIES

TABLE OF CONTENTS

Working Group Roster

NATIONAL INSTITUTES OF HEALTH

National Cancer Institute
National Cancer Advisory Board
Ad Hoc Working Group on Strategic Approaches and Opportunities in Population Science, Epidemiology, and Disparities

CO-CHAIR
Julie R. Palmer, Sc.D.
Karin Grunebaum Professor of Cancer Research
Boston University School of Medicine
Associate Director
Slone Epidemiology Center
Co-Director
BU-BMC Cancer Center
Boston, Massachusetts

CO-CHAIR
Leslie L. Robison, Ph.D.
Chair, Department of Epidemiology and Cancer Control
Associate Director
Cancer Prevention and Control
Comprehensive Cancer Center
Co-Leader, Cancer Control and Survivorship Program
St. Jude Children's Research Hospital
Memphis, Tennessee

MEMBERS

Lucile L. Adams-Campbell, Ph.D.
Professor of Oncology
Associate Director
Minority Health and Health Disparities Research
Senior Associate Dean
Community Outreach & Engagement
Lombardi Comprehensive Cancer Center
Georgetown University Medical Center
Washington, D.C.

Christine Ambrosone, Ph.D.
Professor of Oncology
Senior Vice President, Population Sciences
Chair, Cancer Prevention and Control
Roswell Park Alliance Foundation Endowed Chair in Cancer Prevention
Roswell Park Comprehensive Cancer Center
Buffalo, New York

James R. Cerhan, M.D., Ph.D.
Professor of Epidemiology
Chair
Department of Health Sciences Research
Mayo Clinic College of Medicine and Science
Rochester, Minnesota

Judy E. Garber, M.D., M.P.H.
Director
Center for Cancer Genetics and Prevention
Dana-Farber Cancer Institute
Professor of Medicine
Harvard Medical School
Boston, Massachusetts

Maria Elena Martinez, Ph.D. M.P.H.
Professor
Department of Family Medicine
and Public Health
Associate Director, Population Sciences,
Disparities, and Community Engagement
Moores Cancer Center
University of California, San Diego
La Jolla, California

Electra D. Paskett, Ph.D.
Marion N. Rowley Professor of Cancer
Research
Director, Division of Cancer Prevention
and Control
Department of Internal Medicine
College of Medicine
The Ohio State University
Columbus, Ohio

Bruce D. Rapkin, Ph.D.
Professor
Department of Family and Social Medicine
Head, Division of Community Collaboration
and Implementation Sciences
Department of Epidemiology and Population
Health
Albert Einstein College of Medicine
Bronx, New York

Margaret R. Spitz, M.D., M.P.H.
Professor
Department of Medicine
Dan L. Duncan Cancer Center
Baylor College of Medicine
Houston, Texas

Kate Yeager, Ph.D., R.N.
Research Assistant Professor
Nell Hodgson Woodruff School of Nursing
Emory University
Atlanta, Georgia

Executive Secretary
Deborah M. Winn, Ph.D.
Acting Director
Division of Cancer Prevention
National Cancer Institute
National Institutes of Health

Executive Summary

At its November 2017 meeting, the National Cancer Advisory Board (NCAB) of NCI voted to create an ad hoc Working Group on Strategic Approaches and Opportunities in Population Science, Epidemiology, and Disparities. The Working Group, established in May 2018, was charged by NCI Director Ned Sharpless to first develop recommendations regarding how the observational extramurally supported cancer epidemiology cohort program can be enhanced going forward.

In consultation with NCI, the Working Group identified five questions to focus on that would capture key issues related to the extramurally supported cancer epidemiology cohort program. The group met by teleconference on nine occasions for discussion and background briefings provided by experts (listed in Appendix II). This was followed by an in-person meeting of Working Group members on January 24-25, 2019, and two more teleconferences to discuss drafting the report. The five key questions are outlined below.

Question 1. The role of cohort studies in etiologic and survivorship research in human populations

How can NCI ensure that its cancer epidemiology cohort portfolio has the potential to address questions of the future related to cancer risk, cancer recurrence, cancer survival, and cancer-related long-term health outcomes?

Question 2. Utility of cohorts for addressing cancer health disparities

What is the best way to ensure that the portfolio of extramural cohorts includes cohorts with large numbers of one or more populations that have been understudied and underserved?

Question 3. Study design considerations for extramural cancer epidemiology risk and survivor cohorts

What are the optimal study designs to address cancer risk, recurrence, survival, and long-term health-related outcomes following cancer in human populations?

Question 4. Data sharing and collaboration

How can NCI ensure that the extramural scientists responsible for designing, organizing, and maintaining the cancer epidemiology cohorts remain motivated to continue these time-consuming efforts in this era of rapid sharing of data?

Question 5. Funding models for cohorts

Is the funding mechanism to support cancer epidemiology cohorts optimal? If not, what other models might be better?

Based on careful analysis and the consideration of a range of options, the NCAB Working Group on Strategic Approaches and Opportunities in Population Science, Epidemiology, and Disparities recommends that NCI move to implement the following recommendations. Each recommendation is justified and discussed in detail in the Working Group Question Report Narratives section.

Recommendations for Question 1: The role of cohort studies in etiologic and survivorship research in human populations

1. There will inevitably be circumstances where a cohort design reflects the most scientifically rigorous approach, and generally the most cost-effective approach over the long term, to investigate important existing and emerging topics relating to cancer risk and outcomes. Thus, NCI should invest in providing sufficient infrastructure support for cohorts to conduct or facilitate research that addresses critical scientific gaps, anticipates the scientific questions of the future, and considers societal issues that are deemed to be of high importance with high impact.
2. While capitalizing when possible on existing or planned major cohorts such as *All of Us*^{SM} or NCI's Division of Cancer Epidemiology and Genetics (DCEG) *Connect* Cohort, NCI should continue to support new and existing focused cohort studies to address specific cancer etiology and survivorship questions.
3. NCI should promote or facilitate the use of existing and planned intramural cohorts in order to leverage access of these resources for the broader extramural community to conduct research that will inform determinants of cancer risk and health outcomes after a cancer diagnosis.
4. Cancer survivorship cohorts should be designed to facilitate research that spans the period from diagnosis to long-term survival. This may best be achieved by leveraging data available through clinical trials. In the future, electronic medical records may be an excellent source of data for such cohorts, but at present the quality of data available is not adequate, due to incompleteness, limitations of natural language processing, and limited interoperability across records.
5. When considering the establishment of new survivor cohorts, opportunities to leverage the patient populations available through the NCI-supported cooperative clinical trials groups and the NCI Community Oncology Research Program (NCORP) should be given strong consideration. NCI should support the conduct of pilot studies to determine the feasibility and design for establishing an adult survivor cohort to investigate treatment-related adverse outcomes for cancer patients enrolled and not enrolled in a clinical trial. A challenge that must be addressed is collecting adverse outcome data at the same level of detail for those not enrolled in clinical trials.

6. NCI should promote or facilitate the use of prevention and cancer therapy trials to address etiological and survivorship questions after they have met their primary and secondary endpoints. Whenever possible, this should include planning at the beginning of prevention and therapy trials to follow-up with study participants and collect data useful for addressing both etiologic and survival questions and trial-related scientific questions requiring long-term follow-up. It will be critical to involve observational cohort investigators and prevention and therapy trial researchers in enhancing the capacity of cancer prevention treatment trials to collect more and detailed data that will be useful in evaluating the long-term cancer risks, second primary cancers, and other health events occurring subsequent to the end of the trial. Enhanced informed consent documents will also be required.
7. To facilitate cohort-based research, NCI should support establishment of national infrastructure for ascertainment and follow-up of cancer cases (i.e., the Virtual Pooled Registry's and the Surveillance, Epidemiology, and End Results [SEER] program's ability to fully characterize treatment exposures).
8. More exploration of the opportunities for research on biomarkers of early detection in cohort studies is needed.
9. NCI should support methodological research to evaluate the risks, benefits, and optimum approaches for the return of results to cohort participants.
10. As part of the ongoing peer-review process for continued funding of cohorts, investigators should be asked to justify the need for continued follow-up of the members of the cohort, including the anticipated scientific yield. The study section peer-review procedures should include an assessment of the investigators' justification. When the yield from a cohort is deemed to no longer be justified, then consideration should be given to transitioning a cohort to passive follow-up (i.e., linkage with vital statistics) or termination.
11. There are important opportunities to draw upon the strengths/attributes of cohorts to conduct intervention research by (1) identifying opportunities within existing cohorts for the conduct of intervention-based research that would not compromise the primary objectives being addressed within the cohort, and (2) considering study design and infrastructure requirements for future cohorts to maximize opportunities to integrate or facilitate intervention-based research.

Recommendations for Question 2: Utility of cohorts for addressing cancer health disparities

1. Additional cohorts are required in order to fill existing and future gaps in the NCI cohort portfolio with regard to research on underrepresented populations. The goal is to ensure that detailed study of the determinants of risk and cancer outcomes in these populations can be ascertained with high statistical precision. The Working Group identified the following as of highest priority for additional funded cohorts: Hispanics in the U.S.;

Pacific Islanders; American Indians/Alaska Natives; Blacks; persons of low socioeconomic status; and residents of rural Appalachia.

2. Only one cohort includes a sizable number of Hispanics, Pacific Islanders, and Blacks, and the current age range in that cohort is now 71-99. Two other cohorts include large numbers of Black participants, but they too are aging. While the accumulated resources from cohorts that "age out" can be archived for future work, research on cancer in more recent birth cohorts of participants from underrepresented groups will be required in order to study the effects of new or evolving exposures and social conditions.
3. Support additional biospecimen collection (tumor tissue and blood are the highest priorities) in those existing cohorts that have an appreciable number of participants from a single underrepresented group in an appropriate age range to address scientifically important questions.
4. Encourage risk and survivor cohorts to include questions that permit participants to self-identify as sexual and gender minorities (SGM).
5. Provide investigator-initiated funding (e.g., R01 or P01) to conduct multi-cohort collaborative research addressing compelling scientific questions among minority participants with less common cancers such as head and neck, pancreas, kidney, and myeloma, and on specific subtypes of other cancers.
6. It is not necessary for cohort studies to represent all, or even several, race/ethnicity populations. The major goal is to ensure that currently underrepresented groups be represented in sufficient numbers across the entire NCI cohort portfolio to allow for meaningful within-group analyses. Comparisons across population groups is a secondary goal, which can be accomplished in a variety of ways, including comparisons with other cohorts or the published literature. New cohorts can address the major goal through different approaches, including studies of single population groups and studies of multiple groups that oversample minority groups. Future program announcements should note that cohorts of single populations are acceptable.

Recommendations for Question 3: Study design considerations for extramural cancer epidemiology risk and survivor cohorts

1. Cohorts remain a major investment in cancer epidemiology, but also provide some of the greatest scientific impact, and thus ensuring a balanced portfolio of cancer risk and survivor cohorts is important. Etiology cohorts should consider the current or emerging gaps in research and comprise the appropriate populations (age, birth cohorts, and critical windows of exposure) to address the gaps. New survivor cohorts should address current and emerging research gaps by cancer type and/or treatment.
2. Approaches for leveraging innovative sampling frames to recruit study participants should be utilized to maximize the value of cohorts in addressing scientific questions.
3. There is a need to promote improvements in electronic health records (EHR) systems and other digital technologies to enable them to be better utilized as sampling frames, and for

exposure assessment and ascertainment of outcomes for cancer etiology and survivor studies.

4. NCI should identify possible opportunities for embedding cohorts in intervention trials for primary prevention, screening and treatment. Further, NCI should consider joining with other NIH institutes in creating cohorts that could address both cancer outcomes and other health outcomes.
5. Given limited resources, cohorts should generally derive their study populations from the U.S. and its territories. Nevertheless, there may be circumstances where a study population from another country provides unique opportunities that should be pursued when possible.
6. The Working Group strongly encourages, when scientifically justified, the incorporation of serial data and biologic specimen collection cycles over extended periods of time to reduce measurement error for time-dependent events (e.g., quitting smoking) and to enable a better understanding of the natural history of cancer (e.g., how epigenetic or metabolomic or immunological characteristics change over time and influence cancer risk and outcomes).
7. There is a need to adopt innovative methods for data collection from study participants, when appropriate, that may be more accurate, less burdensome, and economical to administer (e.g., mobile technologies).
8. Consideration should be given to study participant preferences (interview vs. questionnaire), abilities (electronic devices), and environmental context (internet access) for providing their data. Innovative and validated approaches should be utilized to maintain bi-directional engagement of cohort participants with the researcher team.
9. The NCI should support or facilitate methodological research to identify efficient and effective approaches for incorporation of longitudinal specimen and data collection into cohort studies.

Recommendations for Question 4: Data sharing and collaboration

1. Guidelines and/or mandates for data sharing of cohort-based data must take into consideration the investment of time and academic implications for the investigators responsible for establishing and maintaining the cohort. Thus, these investigators should have a defined window of opportunity to pursue their own research interests within the cohort, prior to making the data available to the broader research community.
2. Given the investigator and staff time/effort associated with data sharing/collaborative efforts (initial posting and updating of data, subsequent updating of data and associated documentation, review of concept proposals, preparation of user-friendly data files and associated documentation, and responding to questions) ongoing funding for data sharing will be needed. Supplements have not been an appropriate funding approach for these purposes because of the limited timeline for activities.

3. For existing cohorts, data sharing guidelines should allow for different mechanisms of sharing depending on requirements of the informed consents provided by the cohort participants. In some cases, informed consents may not allow for some types of sharing (e.g., placing individual-level data in a government database that can be accessed by outside investigators without oversight by the cohort investigators), and it may not always be feasible to re-consent participants.
4. For new cohort studies, consent for broad data sharing should be made part of the initial enrollment procedure.
5. Given NCI's investment in data science and the availability of new tools and technology, NCI should invest in the modernization of existing and new cohorts to facilitate sharing, with practices consistent with FAIR (Findable, Accessible, Interoperable, Reusable) principles.
6. If there is an initiative for creation of a centralized data sharing platform(s) for cohort-based data, it should be recognized that, because of the heterogeneity of study designs and associated data elements, it would have to be limited to the set of variables common to the majority of cohorts. This limited set of common variables would likely not meet the needs of many researchers. An alternative could be development of a federated system in which each study has its own data platform, which can be accessed, with appropriate permissions and informed consents, to pull data elements across cohorts.

Recommendations for Question 5: Funding models for cohorts

1. The NCI should continue to use a Cohort Infrastructure Program Announcement for funding infrastructure activities of cancer cohorts. Investigator-initiated hypothesis-driven research based on cohort data would continue to be funded through R grants, P01s, and related mechanisms.
2. Applications for new cohorts should be considered in a special study section, separate from the study section that reviews continuations of cohorts.
3. It may be most effective for the NCI to accept applications for new cohorts only in response to a call for applications, which would occur periodically as needed.
4. Decisions about when to stop funding active follow-up of a given cohort should be made based on the likely productivity and importance of findings that will occur over the next five years.
5. A specific subheading could be added to the cohort infrastructure application to ensure that PIs will give a detailed rationale and justification for continuation of the cohort for another five years.
6. There is a need for further discussion to determine best practices for *whether* and *how* samples should be preserved for future use after funding for a given cohort has ceased, as well as who will make decisions about biospecimen use. At a minimum there is a cost for keeping freezers operating and supporting sample management.

Overview of Observational Cohorts within the NCI Extramural Research Portfolio

The extramural program of the NCI currently supports numerous large observational cancer epidemiology cohorts to study determinants of cancer in human populations and health outcomes among cancer survivors. The cohort study remains one of the strongest sources for evaluating risk of cancer as well as cancer outcomes, including treatment effectiveness. In many settings this is the only practically and/or ethically available source of evidence for clinical and public health decisions and is now being considered by FDA in certain situations to assess efficacy as part of their Real World Framework (U.S. FDA, 2019). Thus, cohort studies have been and will continue to be a fundamental and key resource for generating and evaluating new knowledge across the cancer continuum.

The design, conduct, and maintenance of cohorts addressing cancer etiology and outcomes are supported through a variety of grant funding mechanisms from NCI. An examination of the current NCI portfolio of extramurally-funded cancer cohorts found that:

- NCI supports 20 risk cohorts, for which the outcome studied is incident cancer, and 10 survivor cohorts, with outcomes of long-term morbidity and mortality (listed in Appendix I, Table 1).
- The number of study participants overall and by race and ethnic group across the risk cohorts is shown in Table 2 for risk cohorts and Table 3 for survivor cohorts. Overall, more than 2.8 million U.S. domestic study participants are being followed for the risk cohorts and 86,000 for the survivor cohorts. Of the 2.9 million, 1.4 million are in the Breast Cancer Surveillance Consortium, a breast cancer screened population.
- The most common cancers being investigated in the cohorts (Table 4) reflects the most common cancers in the U.S. population, with breast, prostate, lung, and colon cancer accounting for 78% of all incident cancers across the cohorts to date.
- Most of the cohorts collect biospecimens, including blood, urine, and tumor tissues. Some collect saliva or buccal cells in addition to or in lieu of blood. Two cohorts collected feces. The counts of biospecimens by cohort by type of biospecimen is shown in Table 5.
- The currently funded risk cohort with the longest duration of follow-up is the Nurses' Health Study, which began enrollment in 1979.
- The Southern Community Cohort Study is the most recently established risk cohort in the grant portfolio; it started enrolling study participants in 2002 and completed enrollment in 2009.

Information about the racial and ethnic groups and study participants from understudied demographic groups are described in the section of this report on health disparities. Information about many of these cohort research resources is available in the Cancer Epidemiology Descriptive Cohort Database (CEDCD: https://cedcd.nci.nih.gov/).

NCI funds its extramural cohort portfolio using grants mechanisms. All but one of these grants are managed either by the Epidemiology and Genomics Research Program (EGRP) in the Division of Cancer Control and Population Sciences or by the Division of Cancer Prevention (for several of the cohorts that were generated from cancer prevention trials). The current annual total cost (direct plus indirect costs) as of April 30, 2019 was $34 million for the risk cohorts (Table 1). For the survivor cohorts the amount was $22 million.

Because cancer epidemiology cohorts are viewed as research resources, NCI expects cohort principal investigators to share data and biospecimens with other researchers. The CEDCD provides information about how researchers can access the data and specimens from many of these cohorts. The current NCI announcement (PAR-17-233) for support of core infrastructure and methodological research for cancer epidemiology cohorts includes the following data sharing policy:

> Individuals are required to comply with the instructions for the Resource Sharing Plans as provided in the SF424 (R&R) Application Guide, with the following modification:
>
> NCI expects awardees to propose and implement a robust data sharing plan that details how external investigators gain access to data and biospecimens to ensure that this resource is used widely. Individuals are required to comply with the instructions for the Resource Sharing Plans as provided in the SF424 (R&R) Application Guide, with the following modification:
>
> - All applications, regardless of the amount of direct costs requested for any one year are expected to include a Data Sharing Plan that is compliant with NIH data sharing policies, including the NIH Genomic Data Sharing Policy (https://gds.nih.gov/03policy2.html).
> - CECs are required to maintain a website that details the procedure for requesting and obtaining data for external Investigators as appropriate and consistent with achieving the goals of the program. A summary of the number of data requests, acceptances, and rejections should be provided in annual progress reports to NCI.
> - Awardees are strongly encouraged to deposit individual-level de-identified datasets to NCI's centralized, controlled-access database, called the Cancer Epidemiology Data Repository (CEDR). More information about this resource is available at https://epi.grants.cancer.gov/CEDR.

Most of the risk cohorts are members of the NCI Cohort Consortium (https://epi.grants.cancer.gov/Consortia/cohort.html), which enables investigators to pool cohort data to conduct studies where there are not adequate numbers of participants for investigation of rarer cancers, molecular subtypes of common cancers, rare exposures, and smaller demographic strata among other factors. The Cohort Consortium and the willingness of principal investigators of large cancer epidemiology cohort studies to pool data and analyze DNA from study participants have played an important role in the conduct of genome-wide association studies of numerous types of cancer.

What is not covered in this report. This report does not cover specialized risk cohorts whose study participants may have been ascertained because of a characteristic such being HIV-positive or exposure because they work in a specific industry. It does not include cohorts that NCI may fund in other countries for specific purposes such as to investigate populations with very high risks of particular cancers or unusual exposures. This report does not cover all aspects of cancer survivorship, but rather focuses primarily on the outcomes of mortality and survival, recurrence of the primary cancer, incidence of new primaries, and certain other physical sequelae. It is recognized that many cancer survivor cohorts collect a wider scope of content domains, including adverse sequelae after a cancer diagnosis, psychosocial concerns, health-related quality of life, health behaviors, patterns of care, and the economic impact of cancer. While these other domains are important topics for study and cohorts that include these domains are of high scientific, clinical, and psychosocial value, developing recommendations on these other domains fell outside of the scope of this report.

Working Group Question Report Narratives

Question 1. The role of cohort studies in etiologic and survivorship research in human populations

Background

A robust and scientifically rigorous portfolio of research addressing the etiology and outcomes of the more than 1.7 million new cancer cases diagnosed each year in the U.S. (Siegel et al., 2019) is important to the mission of NCI. Projections of cancer incidence and mortality, taking into consideration changes in the demographics of the U.S. population (Figure 1), emphasize the dynamic landscape and impact of cancer within the population (Rahib et al., 2014). Moreover, with the improvements in cancer detection, diagnosis, and treatment, it is estimated that there will be 18.1 million cancer survivors in the U.S. by 2020 and this is projected to increase by 31%, to 20.3 million, by 2026 (Bluethmann, et al., 2016) (Figure 2).

In addition to the heterogeneity of cancer within and across types, the documented differences in cancer incidence, mortality, and long-term outcomes by factors such as sex, age, race/ethnicity, and socioeconomic status highlight the complexities of cancer research. Observational cancer research has contributed significantly to the understanding of determinants of risk for cancer and their overall contribution to cancer incidence. Major contributors to cancer risk, such as cigarette smoking, excess body weight, alcohol consumption, UV radiation, poor diet, infections, physical inactivity, and genetic predisposition reflect opportunities for prevention and/or early detection. Importantly, beyond established risk factors there exist a myriad of factors (e.g., temporal changes in environmental and lifestyle-related exposures, access to and use of health care, social and economic status, cultural beliefs, physical and mental health, health literacy, cancer-related therapy) that directly influence cancer risk, prevention, treatment, and survivorship. Within the context of performing scientifically rigorous and highly impactful research, consideration of this constellation of influences on cancer risk and outcomes is often best conducted through the establishment of well-characterized cohorts.

Scientific discoveries from cancer epidemiology cohorts help explain patterns of cancer incidence and mortality and inform biological mechanisms of carcinogenesis. Results from cohorts often provide the evidence base for risk assessment and risk prediction, as well as changes in health practice and policy. For example, cohort findings have been extensively used to inform clinical and public health guidelines and identify opportunities for public health and clinical interventions to prevent cancer; to detect cancer at an earlier point when interventions may be more effective; and to ameliorate the consequences of cancer. Public health policies often draw heavily from cohort study findings.

There are several recent large intramural NIH cohort-based initiatives designed to address cancer-related risk and cancer outcomes. The Working Group members were briefed by lead investigators of the NIH's *All of Us*^{SM} Research Program and NCI *Connect* Cohort, both of

which could potentially be a resource for the extramural community to examine the determinants of cancer and outcomes after a cancer diagnosis. Brief summaries of the *All of Us*^{SM} cohort and NCI's *Connect* Cohort are available in Appendix III.

There are many examples of current successes of repurposing cancer-related prevention trials to address questions about determinants of cancer risk such as the NCI-supported PLCO Cancer Screening Trial, SELECT, and the Women's Health Initiative. Some trials unrelated to cancer also have been transformed into cancer epidemiology risk and survivor cohorts; one example is the Atherosclerosis Research in Communities Study (ARIC).

Figure 1A. Incidence projections of selected cancers by 2030 due to demographic changes and the average annual percent changes in incidence rates. Figure 1B. Death projections of specific cancers based on demographic changes and the average annual percent changes in death rates. (Source: Rahib et al., 2014)

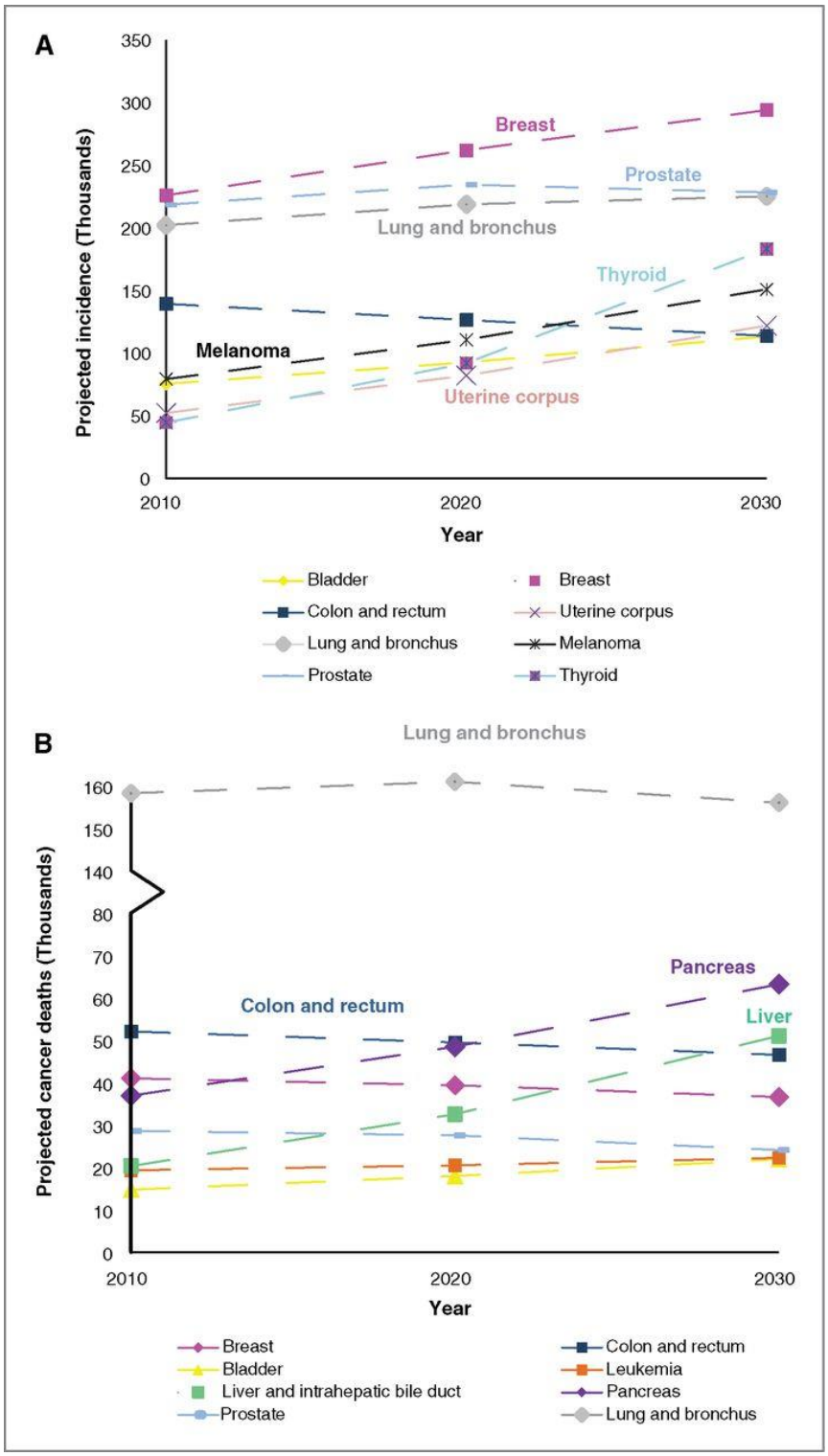
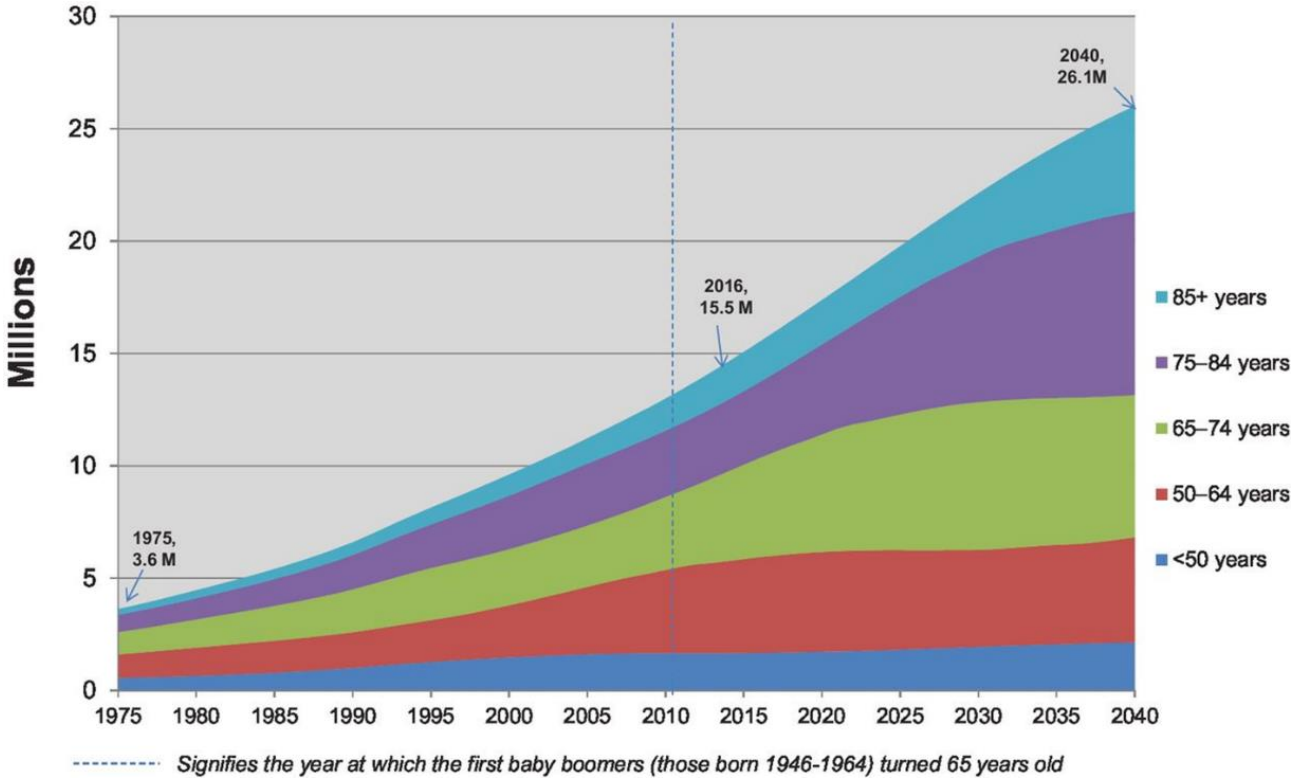


Figure 2. Estimated cancer prevalence by age in the U.S. population from 1975 to 2040. (Source: Bluethmann et al., 2016)



Working Group Assessment

Cohort studies have been an important source of information relating to the current understanding of cancer and should continue to be an integral part of the NCI research program's strategy for addressing its mission relating to etiology, prevention, and survivorship. Within the context of the NCI-funded cohorts there should be ongoing review to ensure diversity across the cohort portfolio relative to study populations (diversity in people, exposures, diseases, geography), cancer types, understanding disparities in incidence and mortality/survival, emerging issues in cancer including the broad spectrum of issues related to cancer survivorship (e.g., treatment and treatment-related risks, healthcare access, and coverage).

There will be an ongoing need for new etiology-related cohorts to address future areas of research, including: (1) high-risk populations for understudied cancers, (2) racial/ethnic and other diversity of populations, (3) new exposures within the population, (4) integration of new technologies, (5) risk assessment and risk prediction, (6) the natural history of progression to cancer through evaluation of changes in biomarkers measured using non-invasive means, (7) opportunities for early detection by identifying biomarkers for early detection of incident

cancers, and (8) opportunities for intervention research. Examples include cancers with increasing incidence and/or mortality (e.g., pancreatic, liver), racial/ethnic differences in incidence and/or mortality, temporal changes in health behaviors or environmental factors associated with cancer risk (e.g., tobacco/nicotine, obesity, UV radiation, infections), how changes in physiological characteristics (hormonal and metabolic profiles) may influence cancer risk, advancing mobile health technologies (e.g., mobile phone applications and wearable sensors), the expanding repertoire of biomarkers for earlier detection, genetic/familial factors, access to care, and early detection.

To date, cohort studies have not been commonly used for research on biomarkers for early detection. Advances in identification of early detection biomarkers, including circulating tumor cells or circulating tumor DNA, offer opportunities for high impact research in prospective cohorts. Most cohorts with biospecimens collected the biospecimens at baseline or a single time during follow-up; few cohorts have serial biospecimen collection on an appreciable number of participants. Study participants who developed an incident cancer within a few years of a blood draw are useful for research on early detection biomarkers but may be limited in numbers. Serial longitudinal collection of biospecimens would greatly increase the power for this type of research.

Existing cohorts reflect important resources and significant investments in cancer-related research. Thus, there needs to be careful consideration of how best to retain the scope and scientific integrity of these valuable resources over time, while maximizing the yield in scientific output. Critical consideration needs to be given to the evaluation of existing NCI-funded cohorts relative to current and future potential for productivity as well as opportunities to implement cost-saving approaches and to expand/enhance the investments that have already been made. When, as part of the peer-review process, cohorts are determined to have limited future productivity, strategies for phased termination and strategies to archive data and biospecimen resources for future use will be important.

Cancer survivor cohorts provide data critical to (1) determine the incidence and severity of treatment-related adverse outcomes, (2) identify populations at highest risk for adverse health and quality of life outcomes, (3) inform development of long-term follow-up guidelines, (4) guide the design and testing of intervention strategies, and (5) inform the dissemination and implementation of efficacious interventions to prevent or ameliorate late effects of cancer treatment. Further support for the growing importance of survivorship cohorts is demonstrated by the FDA's interest in incorporating real world data (highlighted by use of well-designed and conducted observational cohorts) as part of their real-world evidence program (U.S. FDA, 2019).

New survivor cohorts will be required to investigate recurrence, factors contributing to new primary cancers, and survival and mortality among cancer survivors with respect to: (1) the introduction of new anti-cancer agents and diagnostic and therapeutic approaches across the spectrum of cancer treatment types, including radiation therapy sources/techniques

chemotherapeutic agents/regimens, immunotherapy, targeted therapy, and surgery; (2) the expanding repertoire of biomarkers of response to cancer treatment; (3) the long-term outcomes in survivors of understudied cancer site; and, (4) the changing demographics of cancer survivors and increasing duration of survival.

Beyond the current NCI funded portfolio, there are several major intramural initiatives, including the *All of Us*[SM] Research Program and the DCEG Connect Cohort (see Appendix III), that will eventually be shared resources available for the extramural community to utilize in examining the determinants of cancer and outcomes after a cancer diagnosis. These resources should be considered within the context of future needs relative to cancer research questions. While these NIH cohorts have the potential to pursue scientific questions related to cancer, they are not in and of themselves adequate to address the scope and heterogeneity of research questions related to cancer risk and survivorship. In addition, the potential for the extramural research community to add data collection components to these studies is uncertain. Thus, there continues to be a need to support cohorts that address specific scientific questions, utilize special study designs, examine specific exposures, and/or have more focused sets of outcomes with a greater depth of detail about those outcomes. Rarely will one cohort be able to answer or address all scientific questions, which underscores the need for a cohort portfolio.

While some large prevention trials have been repurposed to study risk and survivorship (e.g., the Women's Health Initiative and the Physicians' Health Study), after the trials' primary endpoints have been achieved, the Working Group believes there is an opportunity to more fully exploit both prevention and cancer treatment trials for risk and survivorship research. Ideally, this would involve including such plans in the original trial protocols to facilitate activities such as participant reported data, re-consenting, and data linkage activities.

Patient engagement and issues such as return of results are important considerations in long-term prospective studies. Study participants are increasingly being considered as partners in the research enterprise and two-way interaction between researchers and study participants is increasingly expected by study participants. Cohort investigator teams typically maintain a close relationship with study participants to enhance retention in the cohort over decades and recognize their contributions as study participants. Small incentives and recognitions, as well as mailing newsletters describing findings of the study in an aggregate, lay-friendly form, are common. Return of results is a challenging issue in cohort studies because (1) many of the investigations that use biospecimens take place years after the blood draw or other specimen was obtained; (2) many or most assays are conducted in research laboratories that are not CLIA-certified; and (3) additional personnel effort would be required for return of results in a clear, informative, and ethical manner.

Several cohorts have undertaken randomized or other types of behavioral or clinical trials within the cohorts. Examples of such trials are (1) the randomized controlled trial within the Black Women's Health Study to assess efficacy of an insomnia prevention tool and (2) the randomized

controlled trial conducted within the Childhood Cancer Survivor Study cohort to increase uptake of screening mammography among high-risk women treated for childhood cancer with chest radiotherapy. Consideration should be given to the impact, feasibility, and types of trial that could be considered to imbed within ongoing prospective studies. For example, it may be most feasible for low-risk behavioral interventions to be incorporated into a cohort study. Examples of such interventions could be a randomized trial of an intervention to help cancer survivors stop smoking. While there is great potential for intervention-based research within cohorts, it is important to recognize that investigators may be reluctant to incorporate interventions that (1) have the potential to compromise the ability to address the primary objectives of the cohort, or (2) substantially increase respondent burden and decrease cohort retention rates. Some of the concerns regarding impact on outcomes may be addressed statistically by considering the intervention and control conditions as "exposures." In addition, study participants might find opportunities to participate in intervention trials meaningful, which might help improve retention and engagement. This topic deserves additional consideration.

Recommendations and Opportunities for Enhancement

1. There will inevitably be circumstances where a cohort design reflects the most scientifically rigorous approach, and generally the most cost-effective approach over the long term, to investigate important existing and emerging topics relating to cancer risk and outcomes. Thus, NCI should invest in providing sufficient infrastructure support for cohorts to conduct or facilitate research that addresses critical scientific gaps, anticipates the scientific questions of the future, and considers societal issues that are deemed to be of high importance with high impact.
2. While capitalizing when possible on existing or planned major cohorts such as *All of Us*^{SM} or NCI's *Connect* Cohort, NCI should continue to support new and existing focused cohort studies to address specific cancer etiology and survivorship questions.
3. NCI should promote or facilitate the use of existing and planned intramural cohorts in order to leverage access of these resources for the broader extramural community to conduct research that will inform determinants of cancer risk and health outcomes after a cancer diagnosis.
4. Cancer survivorship cohorts should be designed to facilitate research that spans the period from diagnosis to long-term survival. This may best be achieved by leveraging data available through clinical trials. In the future, electronic medical records may be an excellent source of data for such cohorts, but at present the quality of data available is not adequate, due to incompleteness, limitations of natural language processing, and limited interoperability across records.
5. When considering the establishment of new survivor cohorts, opportunities to leverage the patient populations available through the NCI-supported cooperative clinical trials groups and NCORP should be given strong consideration. NCI should support the conduct of pilot studies to determine the feasibility and design for establishing an adult

survivor cohort to investigate treatment-related adverse outcomes for cancer patients enrolled and not enrolled in a clinical trial. A challenge that must be addressed is collecting adverse outcome data at the same level of detail for those not enrolled in clinical trials.

6. NCI should promote or facilitate the use of prevention and cancer therapy trials to address etiological and survivorship questions after they have met their primary and secondary endpoints. Whenever possible, this should include planning at the beginning of prevention and therapy trials to follow-up with study participants and collect data useful for addressing both etiologic and survival questions and trial-related scientific questions requiring long-term follow-up. It will be critical to involve observational cohort investigators and prevention and therapy trial researchers in enhancing the capacity of cancer prevention treatment trials to collect more and detailed data that will be useful in evaluating the long-term cancer risks, second primary cancers, and other health events occurring subsequent to the end of the trial. Enhanced informed consent documents will also be required.
7. To facilitate cohort-based research, NCI should support establishment of national infrastructure for ascertainment and follow-up of cancer cases (i.e., the Virtual Pooled Registry's and the SEER program's ability to fully characterize treatment exposures).
8. More exploration of the opportunities for research on biomarkers of early detection in cohort studies is needed.
9. NCI should support methodological research to evaluate the risks, benefits, and optimum approaches for the return of results to cohort participants.
10. As part of the ongoing peer-review process for continued funding of cohorts, investigators should be asked to justify the need for continued follow-up of the members of the cohort, including the anticipated scientific yield. The study section peer-review procedures should include an assessment of the investigators' justification. When the yield from a cohort is deemed to no longer be justified, then consideration should be given to transitioning a cohort to passive follow-up (i.e., linkage with vital statistics) or termination.
11. There are important opportunities to draw upon the strengths/attributes of cohorts to conduct intervention research by (1) identifying opportunities within existing cohorts for the conduct of intervention-based research that would not compromise the primary objectives being addressed within the cohort, and (2) considering study design and infrastructure requirements for future cohorts to maximize opportunities to integrate or facilitate intervention-based research.

Question 2. Utility of cohorts for addressing cancer health disparities

Background

Incidence data from the CDC-funded and NCI-funded population-based cancer registry programs provide the best resource for identifying disparities in risk and survivorship, with data available on incidence and mortality within U.S. populations defined by race/ethnicity. The most well-established and striking disparities noted are the higher incidence and/or mortality of many cancers for Black men and women relative to White men and women. In the most recent U.S. cancer incidence data (2008-2014), for all cancer sites combined, Black men had the highest incidence rates compared with other racial groups, and Black men and Black women had the highest death rates compared with other racial groups (Cronin et al., 2019).

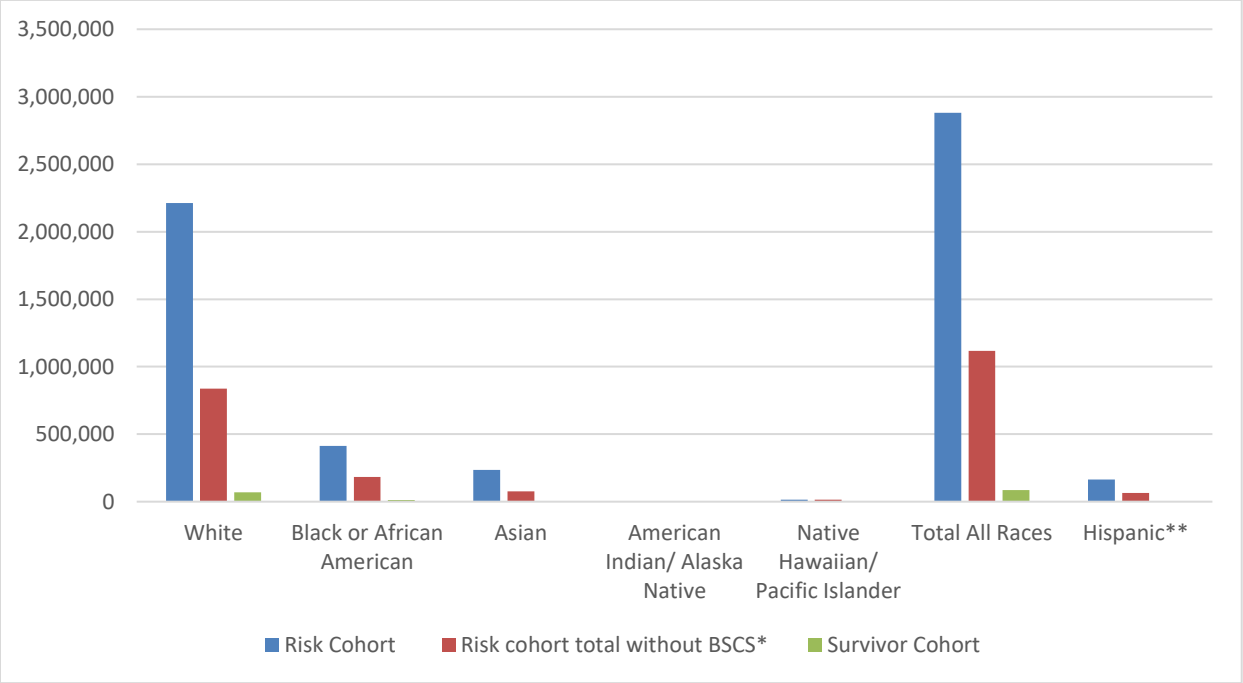
Hispanic/Latinos are among the fastest growing minority group in the U.S., expected to account for 35% of the U.S. population by 2050. Compared to populations of European ancestry, Hispanics, as a single group, have lower rates of the most common cancers and higher rates of some of the less common malignancies that have an infectious etiology. Hispanic ethnicity comprises an admixed population (European, Indigenous American, and African ancestry). This ethnic group is also highly heterogeneous in regard to nativity, culture, and socioeconomic status. Opportunities exist to disentangle this complex heterogeneity when assessing cancer risk and survival in Hispanics. To do this, investment is needed in new cohorts that include a sizeable number of Hispanics; this group has been under-represented in cancer cohorts to date. It will be important to ensure that in addition to including Hispanics in the Southwest, who have significant Indigenous American ancestry, these cohorts also include Hispanics with higher levels of African ancestry, such as individuals from Puerto Rico and the Dominican Republic, as ancestry may play an important role in conflicting data regarding risk and survival.

Important cancer health disparities also exist for subgroups of the U.S. population defined not by race/ethnicity but by other population level factors such as socioeconomic status (poverty), urbanicity (rural), geography (Appalachia, Mississippi Delta, and other regions), and sexual orientation or gender identity (e.g., gay, lesbian, bisexual, transgender, queer). Individuals often belong to more than one of these groups.

To date, the overwhelming majority of participants in cohorts funded by NCI are non-Hispanic Whites (Figure 3). Only three cohort studies (MEC: Multi-ethnic Cohort, BWHS: Black Women's Health Study, and SCCS: Southern Community Cohort Study) have enough Black participants to permit analysis of risk factors for cancer in Blacks, and this is only for the most common cancers – breast, prostate, colorectal, and lung. The MEC provides data for risk analyses for cancer participants of Japanese, Hawaiian, and Latin American descent. None of the NCI cohorts have appreciable numbers of American Indians or Alaska Native individuals. The four international cohorts supported by EGRP follow residents of China and Singapore; those

cohorts are valuable for several reasons, including wider distributions of exposures of possible interest, but they do not contribute to an understanding of cancer health disparities in the U.S.

Figure 3. Race and ethnicity of participants in NCI extramural cohorts



* BSCS (Breast Cancer Screening Consortium), the largest cohort, is removed from this total because breast cancer is the only outcome and limited exposure data were obtained.
** Hispanic ethnicity ascertained separately from race.

Working Group Assessment

The Working Group noted that mortality rates are higher in Blacks than Whites for most cancers, regardless of incidence patterns. Research to date has considered differences in tumor biology, germline genetics, access to care, geospatial factors, and comorbidities. However, reasons for the mortality disparities have not been established, even for the most common cancers: breast and prostate. The same is true for disparities in cancer incidence: although there has been substantial research on both breast and prostate cancer in Blacks in recent years, the reasons for the higher incidence of estrogen receptor negative breast cancer in Black women and aggressive prostate cancer in Black men are not fully understood. All other cancers have received far less attention, in part due to the limited data available from existing cohorts. In particular, the Working Group noted a need for etiologic research on cancers of the pancreas, liver, head and neck, kidney, multiple myeloma, ovary, endometrium, colon and rectum, and lung in Black men and women due to either more aggressive tumors or higher incidence of these cancers in Blacks than Whites.

With regard to mortality, the Working Group noted that both melanoma and ovarian cancer are more lethal in Blacks than Whites, even though incidence is lower.

Survivorship statistics and projections typically have not included a breakdown by race, ethnicity, and geographic location. Research on disparities in cancer survivors has been restricted by the practice of defining eligibility for survivor cohorts as a set time after diagnosis (e.g., five years after diagnosis). Those who die before the first cut-off (often those from disadvantaged groups) are missed by this approach.

The Working Group discussed approaches to increasing the number of cohort participants from underrepresented groups such as Hispanics, Blacks, Native Americans, rural, low income, and sexual/gender minorities. The consensus was that the preferred approach is to focus enrollment on attaining a large number of participants in one or more underrepresented groups, rather than trying to represent all groups at levels proportional to the entire U.S. population. This approach will maximize statistical power for informative analyses related to the etiology of cancer and cancer survivorship in understudied populations. It will also permit the disentanglement of race/ethnicity with socioeconomic status and various cultural factors. The fact that not all minorities are underserved or poor, and that not all poor people are minorities is not universally understood. While there are differences in incidence rates across race/ethnicity, not everyone in the group with the highest incidence develops the disease. To understand the etiology and, perhaps, identify modifiable factors, it is necessary to determine what the individuals who are doing well have in common. For the purposes of reducing cancer health disparities, differences *within* groups (e.g., groups defined by race/ethnicity) are more informative than differences *between* them. Understanding differences within a single group (e.g., why are some Black men developing head and neck cancer while others are not) is more likely to lead to elucidation of etiology and identification of risk and preventive factors of relevance to the population. The Working Group emphasized that reviewers should be educated to understand that cohorts focused on specific subpopulations do not necessarily require inclusion of a comparison group of majority participants (e.g., non-Hispanic Whites or individuals of a high socio-economic status).

Concern was raised about whether recently initiated or planned cohorts will be successful in enrolling large numbers of participants from underrepresented groups in the U.S. It was noted that the fastest enrollment always occurs among Whites, especially those in mid-to-high income levels. If enrollment of those groups is not halted when a predetermined number has been reached, it is unlikely that the goal of having informative data from other subgroups will be attained.

There was discussion of difficulties in studying health outcomes in rural populations. The health implications of living in a rural area differ across different parts of the U.S. It may be best to

design studies to address specific rural populations (e.g., Appalachia, Mississippi Delta, Central Plains, Western).

There was discussion of whether conducting collaborative research with cohorts supported by the National Heart, Lung and Blood Institute (e.g., the Jackson Heart Study and the Hispanic Community Health Study/Study of Latinos) would be a short-term solution to understanding the reasons for racial/ethnic disparities in cancer. It was agreed that those studies, although they have deep phenotype data, are generally too small to be very useful for most cancer etiology research.

The group discussed NCI funding of cohorts in other countries (e.g., China and Singapore), given that meaningful minority data would be best gleaned from U.S. residents. There was agreement that that those studies are not intended to replace the need for studies of U.S. residents of the same ancestry; rather, they are valuable as resources to assess for example the higher incidence of certain cancers or exposures for which variation may be greater in another country than in the U.S. However, participants in foreign cohorts should not be counted as "minority" participants in assessments of the NCI cohort portfolio; instead, they belong in a separate category, such as "non-U.S."

Recommendations and Opportunities for Enhancement

1. Additional cohorts are required in order to fill existing and future gaps in the NCI cohort portfolio with regard to research on underrepresented populations. The goal is to ensure that detailed study of the determinants of risk and cancer outcomes in these populations can be ascertained with high statistical precision. The Working Group identified the following as of highest priority for additional funded cohorts: Hispanics in the U.S.; Pacific Islanders; American Indians/Alaska Natives; Blacks; persons of low socioeconomic status; and residents of rural Appalachia.
2. Only one cohort includes a sizable number of Hispanics, Pacific Islanders, and Blacks, and the current age range in that cohort is now 71-99. Two other cohorts include large numbers of Black participants, but they too are aging. While the accumulated resources from cohorts that "age out" can be archived for future work, research on cancer in more recent birth cohorts of participants from underrepresented groups will be required in order to study the effects of new or evolving exposures and social conditions.
3. Support additional biospecimen collection (tumor tissue and blood are the highest priorities) in those existing cohorts that have an appreciable number of participants from a single underrepresented group in an appropriate age range to address scientifically important questions.
4. Encourage risk and survivor cohorts to include questions that permit participants to self-identify as sexual and gender minorities (SGM).

5. Provide investigator-initiated funding (e.g., R01 or P01) to conduct multi-cohort collaborative research addressing compelling scientific questions among minority participants with less common cancers such as head and neck, pancreas, kidney, and myeloma, and on specific subtypes of other cancers.
6. It is not necessary for cohort studies to represent all, or even several, race/ethnicity populations. The major goal is to ensure that currently underrepresented groups be represented in sufficient numbers across the entire NCI cohort portfolio to allow for meaningful within-group analyses. Comparisons across population groups is a secondary goal, which can be accomplished in a variety of ways, including comparisons with other cohorts or the published literature. New cohorts can address the major goal through different approaches, including studies of single population groups and studies of multiple groups that oversample minority groups. Future program announcements should note that cohorts of single populations are acceptable.

Question 3. Study design considerations for extramural cancer epidemiology risk and survivor cohorts

Background

In the design of a cohort, a few key principles need to be considered, including (1) the often very different needs of etiology, early detection and survivorship cohorts; (2) the need for a range of age and/or birth cohorts; (3) the need to account for specifics of the target population(s) under study; and (4) the need for varying duration of follow-up driven by types of outcomes of interest and event rates. The optimal cohort study design will vary depending on the types of scientific questions that the cohort is meant to address and the target population to which one wants to generalize the results. It is generally not the case that a single cohort can optimally address all questions of interest across the cancer continuum and in all populations of interest, thus requiring a portfolio of cohorts.

Cohort studies of cancer will generally need large sample sizes, large biobanks, and long-term follow-up (defined here as substantially more than five years). Cohorts need to be designed, managed, and funded to optimize multiple and often unanticipated uses (e.g., to incorporate new technologies and new biology that lead to hypotheses that could not have been articulated at the launch of the study). For some questions, particularly involving rare cancers, other study designs such as case-control and family studies will be most appropriate. While there are outstanding examples of joint investment in randomized controlled trials through repurposing of trials to cohorts, highlighted by the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial , the Women's Health Initiative (a primary prevention trial), and CALGB 89803 (Alliance) Trial (a treatment trial), this type of joint partnership has not been extensively utilized by NCI. Furthermore, clinical trials embedded in observational cohort studies—where efficiencies in

patient recruitment based on risk factor profile or genetic background, and detailed clinical information, can be effectively capitalized upon—remain largely untapped.

Working Group Assessment

Studies across the cancer continuum: While, conceptually, cohort studies can simultaneously address questions on cancer etiology, early detection, prognosis, and survivorship, many practical considerations limit this approach. Etiology studies generally require a range of historical and current (at time of enrollment) exposures, long-term follow-up over decades, and large sample sizes (driven by the incidence of specific cancer types or cancer-related outcomes). Linkage to population-based cancer registries (e.g., the NCI SEER program) facilitates efficient conduct of etiology cohort studies but could pose a challenge for predictive, prognostic, or survivor outcome-based studies.

The optimal age at enrollment will vary, but critical health events such as early life exposures that affect cancer risk may be missed when the minimum age at baseline is 40 or above. Study participants may be able to provide their history and recollections of events in their lives up to the baseline; however, memory, measurement error, and other problems can be a concern. Cohorts with younger age at enrollment will need updated exposure assessment and a longer follow-up, with attention to concerns about attrition and loss to follow-up. Birth cohort effects (which cross age and period effects) need to be considered in managing a portfolio of etiology cohorts, as exposure patterns and new exposures at critical times in the life-course (in utero, childhood, adolescence, etc.), as well as medical practice, are continually changing.

Studies of early detection are optimized by having longitudinal samples, allowing evaluation of sensitivity and specificity of biomarkers for cancer detection in populations where tests are most likely to be clinically implemented, which will vary for cancers based on age at onset. Risk factor data can be helpful in this setting, particularly for identifying higher-risk populations.

While studies of cancer prognosis and survivorship can be conducted using a variety of study designs (retrospective and/or prospective) and rely on a variety of data sources (registries, self-report, medical records), they will often optimally need detailed clinical and treatment data at diagnosis, biospecimens collected at diagnosis and around treatment time points, access to pathology tissue, and outcome data that includes disease progression/relapse, acute and long-term treatment toxicities, and re-treatment among other clinical factors. For these factors, population-based cancer registries are often not able to provide these types of data in sufficient detail and closer interaction with health care systems is often required to obtain required data. Also, rapid enrollment of cases at or shortly after diagnosis can be critical for many cancer outcomes studies, in order to obtain pre-treatment biologic samples and to enroll cases with early events that may represent a distinct biology. While clinic-based studies may have significant advantages, they generally lack a population perspective, are susceptible to other selection biases, and may have limited generalizability. While birth cohort effects (and correlated

age/period effects) can be critical for etiology studies, in cancer outcomes diagnostic and treatment eras are usually more critical for providing clinically actionable results.

Target populations. A single cohort study is not likely to be able to meet the needs of all target populations of interest to addressing the cancer burden in the U.S., again requiring a diverse portfolio that addresses the key issues relating to cancer etiology and survivorship (see background to Question 1). Characteristics of study populations that are generally most closely associated with cancer-related health disparities were previously discussed in Question 2 and have implications for study design. Recruitment, implementation, and follow-up approaches and efficiencies of running a cohort will also be driven by specifics of the target population. For example, cohort studies in marginalized, low socioeconomic status, or highly mobile populations provide specific challenges for follow-up that would need to be considered in the design. To build efficiencies, some cohorts have used sampling frames that were designed to include persons who were expected to have good retention, such as nurses and teachers, but this may limit generalizability for some research outcomes. Study populations from other countries may provide unique scientific opportunities, have very high cancer risks or cancer outcome concerns, or unusual exposures and offer an important reason for study, but relevance to the U.S. population may be more limited for some issues. Comparisons across different geographical areas and populations can be valuable. The study of cancer risk differentials across the world, which may vary by as much as 50-fold, has provided valuable insights into cancer etiology.

Design. A key strength of the cohort study design is that enrollment, exposure assessment, and biospecimen banking occur before the outcome(s) of interest. Because of the latency of effect by many exposures, long follow-up of participants is required. This almost universally requires follow-up longer than five years. Indeed, early follow-up is often excluded in etiology cohorts due to concerns of subclinical disease leading to reverse causality. While etiology cohorts generally require the longest follow-up to maximize their potential and maximize the investment (and therefore allow assessment of latency, which is important in designing prevention approaches), even cohorts for early detection and cancer outcomes will have their maximum potential realized with longer follow-up periods. This requirement for extended follow-up has implications for grant mechanisms that support cohorts.

Outcomes. A major strength of cohorts is that they can be used to examine multiple health outcomes. Outcomes studied in cancer etiology cohorts are usually cancer incidence, non-malignant health conditions ascertained by questionnaire with or without verification in health records, health conditions ascertainable from linkage to claims data (e.g., Medicare and Medicaid files), and mortality. Some cohorts link to electronic medical records.

Outcomes studied in survivor cohorts require collection of treatment exposures as well as information on survival, disease progress/recurrence, re-treatment, treatment toxicity, second primary cancers, and long-term health events ascertainable by questionnaire and electronic health records, or other record linkages. Recurrence of cancer is very challenging to collect from health

records, although the SEER program and other initiatives are exploring innovative ways to capture and classify this information. Access to diagnostic (and relapsed) tumor tissue has grown in importance for molecular epidemiology studies to characterize etiologic heterogeneity by tumor phenotype, as well as to understand molecular pathways and response to therapy. The SEER program is developing approaches to maximize its potential as a sampling frame for survivor studies, but it will need to address the lack of detailed information on treatment exposures and delays in identification and enrollment that can hamper prospective enrollment even with rapid reporting.

Evolving opportunities. Evolving biologic knowledge (e.g., in precursor conditions and interest in association with other exposures such as the microbiome), require new types of biologic specimen collection, validation of processing methods, storage, and/or timing of specimen collection, which can be very expensive. Many innovative laboratory assays need to be well validated before they can be deployed in cohort studies involving tens of thousands of people. It is often not clear when the technology and costs have reached this point to enable those technologies to be applied. Technologies enabling the digital revolution in health care offer new opportunities for cohorts in terms of the type and amount of data collection, linkage, and follow-up. As EHR systems mature and become more interoperable, they will also provide new efficiencies and opportunities. Bringing multiple data sources together in a "big data framework" offers new opportunities. Finally, as health care becomes more "consumer driven," cohort studies will need to evolve and embrace this opportunity.

Recommendations and Opportunities for Enhancement

1. Cohorts remain a major investment in cancer epidemiology, but also provide some of the greatest scientific impact, and thus ensuring a balanced portfolio of cancer risk and survivor cohorts is important. Etiology cohorts should consider the current or emerging gaps in research and comprise the appropriate populations (age, birth cohorts, and critical windows of exposure) to address the gaps. New survivor cohorts should address current and emerging research gaps by cancer type and/or treatment.
2. Approaches for leveraging innovative sampling frames to recruit study participants should be utilized to maximize the value of cohorts in addressing scientific questions.
3. There is a need to promote improvements in EHR systems and other digital technologies to enable them to be better utilized as sampling frames, and for exposure assessment and ascertainment of outcomes for cancer etiology and survivor studies.
4. NCI should identify possible opportunities for embedding cohorts in intervention trials for primary prevention, screening and treatment. Further, NCI should consider joining with other NIH institutes in creating cohorts that could address both cancer outcomes and other health outcomes.
5. Given limited resources, cohorts should generally derive their study populations from the U.S. and its territories. Nevertheless, there may be circumstances where a study

population from another country provides unique opportunities that should be pursued when possible.

6. The Working Group strongly encourages, when scientifically justified, the incorporation of serial data and biologic specimen collection cycles over extended periods of time to reduce measurement error for time-dependent events (e.g., quitting smoking) and to enable a better understanding of the natural history of cancer (e.g., how epigenetic or metabolomic or immunological characteristics change over time and influence cancer risk and outcomes).
7. There is a need to adopt innovative methods for data collection from study participants, when appropriate, that may be more accurate, less burdensome, and economical to administer (e.g., mobile technologies).
8. Consideration should be given to study participant preferences (interview vs. questionnaire), abilities (electronic devices), and environmental context (internet access) for providing their data. Innovative and validated approaches should be utilized to maintain bi-directional engagement of cohort participants with the researcher team.
9. The NCI should support or facilitate methodological research to identify efficient and effective approaches for incorporation of longitudinal specimen and data collection into cohort studies.

Question 4. Data sharing and collaboration

Background

Beginning in 2003, the NIH Data Sharing Policy has applied to all grants with NIH funding of $500,000 or more in direct costs in any one year, which, with few exceptions, includes all NCI-funded cohorts. However, for cohorts funded under the Cohort Infrastructure PAR, the Data Sharing Policy applies regardless of funding level, with inclusion of a data sharing plan compliant with the policy or an explanation of why data sharing is not possible (the language on data sharing from the PAR is provided in the Overview section of this report). The precise content of the data sharing plan can vary depending on the data being collected, but generally is expected to include: the mode of data sharing, a schedule for data sharing, the format of the final dataset, description of the documentation to be provided, analytical tools to be provided, and the need for a data sharing agreement. Data submitted to an NIH-supported controlled access repository (e.g., dbGaP) for sharing is encouraged. Data sharing via other repositories is possible if it provides similar accessibility.

In recent years there has been an expanding emphasis and discussion regarding data sharing as a strategy to accelerate discoveries. For example, a set of principles, referred to as the FAIR principles, has been developed that describes a set of properties that data should have to make

them available and sharable with others (Wilkinson, 2016). NIH data sharing policies are continuing to evolve. To date, much of the discussion and infrastructure development for organized data sharing has been focused on genetic and genomic data. More recently, there has been discussion of approaches for sharing of non-genomic data. Funding supplements have been made available to help researchers engage in data sharing activities.

The consent forms used by most cohorts included provisions that allowed for broad data sharing with other investigators, but have varying data use limitations related to what was contained in the consent forms and other factors related to, for example, how data can be used, who can access it, and how patient privacy is managed. Most cohorts extensively and frequently share their data collaboratively with other investigators and also pool data with other cohorts through the NCI Cohort Consortium.

Working Group Assessment

It is generally agreed that data sharing and/or collaboration are an important objective for maximizing the yield from scientific research and that trainees can be a major beneficiary of data sharing activities. The major questions to be addressed for data sharing/collaboration involving cohorts relate to which strategies are most appropriate and effective to promote or facilitate research within the broader research communities. The Working Group noted that cohort data currently are being shared with investigators at multiple levels ranging from individual researchers conducting specific analyses to multi-cohort collaborations. The scope and type of sharing activities are also broad in nature and include genetic/biomarker, clinical, and questionnaire data, as well as biological samples.

While, philosophically, data sharing is positively viewed, it is important to recognize the practical considerations involved. First, cohort investigators make a significant commitment of time to participate in the design and conduct of a cohort, which will directly impact their academic careers based upon the level of success and productivity derived from the cohort. Second, commitments of investigator and staff time and effort are associated with data sharing/collaborative efforts, including review of concept proposals, preparation of user-friendly data files and associated documentation as well as responses to ongoing questions. Third, cohort investigators have an ongoing responsibility to the study participants and the wider community to safeguard the integrity of the cohort through stewardship, oversight, and curation of the data. The group noted that cohort investigator engagement is frequently required to facilitate training and provide expertise in understanding the complexities of the underlying data structures in order to conduct valid analyses of the epidemiological, clinical and genomic data from cohorts that often have repeated measures over time. Because of the complexity and the concern about possible misinterpretations, some existing NCI-funded cohorts are developing their own customized data sharing platforms for making data available to the broader research community. While generally similar in overall approach and features, the unique aspects of individual cohort study designs, in combination with the cohort-specific data elements, mandates that a system

designed to allow optimal data access must be tailored specifically for an individual cohort. Given this need to individually tailor data access systems, a federated data system may be the optimal strategy for sharing of data among multiple cohorts.

Recommendations and Opportunities for Enhancement

1. Guidelines and/or mandates for data sharing of cohort-based data must take into consideration the investment of time and academic implications for the investigators responsible for establishing and maintaining the cohort. Thus, these investigators should have a defined window of opportunity to pursue their own research interests within the cohort, prior to making the data available to the broader research community.
2. Given the investigator and staff time/effort associated with data sharing/collaborative efforts (initial posting and updating of data, subsequent updating of data and associated documentation, review of concept proposals, preparation of user-friendly data files and associated documentation, and responding to questions) ongoing funding for data sharing will be needed. Supplements have not been an appropriate funding approach for these purposes because of the limited timeline for activities.
3. For existing cohorts, data sharing guidelines should allow for different mechanisms of sharing depending on requirements of the informed consents provided by the cohort participants. In some cases, informed consents may not allow for some types of sharing (e.g., placing individual-level data in a government database that can be accessed by outside investigators without oversight by the cohort investigators), and it may not always be feasible to re-consent participants.
4. For new cohort studies, consent for broad data sharing should be made part of the initial enrollment procedure.
5. Given NCI's investment in data science and the availability of new tools and technology, NCI should invest in the modernization of existing and new cohorts to facilitate sharing, with practices consistent with FAIR principles.
6. If there is an initiative for creation of a centralized data sharing platform(s) for cohort-based data, it should be recognized that, because of the heterogeneity of study designs and associated data elements, it would have to be limited to the set of variables common to the majority of cohorts. This limited set of common variables would likely not meet the needs of many researchers. An alternative could be development of a federated system in which each study has its own data platform, which can be accessed, with appropriate permissions and informed consents, to pull data elements across cohorts.

Question 5. Funding models for cohorts

Background

Several different mechanisms have been used for funding large cancer cohorts. Until 2008, most NCI-funded cohort studies were funded under the R01 or P01 mechanisms. The Cohort Infrastructure PAR separated infrastructure support of the cohort from support for the conduct of research, other than small methodologic research projects. Many of the cohorts are funded through this mechanism. The mechanism changed from the R01 (investigator-initiated grant) to the UM1 and then to the U01 (cooperative agreement mechanism), and this funding mechanism permitted the infrastructure grants to be reviewed in a special study section. Under the current PAR (PAR-17-233), applicants are allowed 30 pages instead of the standard 12 pages for an R01. Applicants are asked to describe the rationale for the cohort; document feasibility of enrollment, follow-up success, and vital status of participants; describe plans for future data collection; and describe the anticipated uses for scientific research. By contrast, an R01 is limited to 12 pages and must use most of the space to justify the science and provide preliminary scientific results, detailed methods related specifically to the science, and an extensive data analysis plan, leaving little room to discuss the actual cohort study itself.

Initially there was no cap for cohort infrastructure grants; subsequent PARs have instituted direct cost budget caps with a current cap of $1.25 million per year. Grants are typically funded for five years.

One active cohort grant that includes infrastructure support (the Childhood Cancer Survivor Study) is funded under the U24 resource grant. The U24 application describes the resource, how the resource is being maintained, enhanced, and utilized for research, and how the resource is scientifically sound. A key component of a U24 grant is allowing the cohort data to be utilized by a broad community of researchers. To this end, the U24 grant may fund biostatistical support to perform analyses at a statistical center as part of the overall resource.

Currently, there is not a consensus at the NCI as to the best mechanism of support for cohort studies. Some have questioned whether the current approach of separating infrastructure support from research funding improves scientific productivity or efficiency in funding. Separation of the research proposals from the infrastructure proposals may make it difficult to assess if the appropriate data and biospecimens are being collected to address the anticipated research activities. Under the Cohort Infrastructure PAR mechanism, the proposed research agendas are not detailed and reviewers may therefore have difficulty assessing the appropriateness of the infrastructure proposal. Some at NCI would prefer to see a hypothesis-driven grant mechanism for new cohorts to ensure that the appropriate infrastructure is in place to address the scientific questions. An additional challenge to the current mechanism is determining whether budgeted items are more appropriate for a research grant rather than basic maintenance support. For

example, the NCI has not permitted investigators to included genetic/genomic assays in an infrastructure grant, based on the premise that reviewers would not have enough information to appropriately review proposed methods for genetic/genomic tests. A hypothesis-driven grant mechanism would also allow for support of scientific effort from co-investigators; at present very little effort is allowed for investigators, who are included only to the extent that they are required for oversight and management of various infrastructure activities.

Given the demands on research funding, an additional concern is how new cohorts can be formed to address research gaps without sunsetting existing cohorts whose populations are close to their lifespan.

Working Group Assessment

The Working Group's discussion was focused on funding mechanisms, inception of new cohorts, and "sunsetting" of cohorts.

Working group members raised the following points in reference to the idea of changing the cohort funding mechanism back to investigator-initiated R01s.

1. It may become very difficult for new cohorts to be funded (or re-funded) because hypothesis testing would have to be completed during the initial five-year period, while data collection and follow-up are still in early stages.
2. The 12-page limitation for R01 applications is too restrictive and would be inadequate to describe the cohort (enrollment, data collection, tracking, and collection of biospecimens) in addition to specific hypotheses (rationale, preliminary data, detailed methods/quality control procedures specific to the hypothesis, statistical plan, etc.).
3. The strength of a cohort study lies in the continuity of longitudinal data collection and follow-up. Relying on R01s to support the cohort infrastructure poses a greater risk of loss of that continuity.
4. A strength of the current infrastructure PAR approach is that it appears to have helped DCCPS make decisions on continued funding of existing cohorts and funding for new cohorts in the context of considering the entire portfolio of cancer cohorts. The PAR, along with use of the Awaiting Receipt of Applications (ARA) process, has led to greater clarity about potential for funding.
5. The U01 mechanism implies an institutional partnership, which Working Group members agreed is desirable.
6. The lack of clarity about which aspects of data/biospecimen collection are funded through a cohort infrastructure grant versus through a science R01 is a problem. For example, investigators submitting hypothesis-driven R01s have been told they cannot include obtaining tumor tissue in the budget if their cohort infrastructure grant supports

tumor tissue collection. In this instance, problems would arise if the infrastructure grant has fewer remaining years than projected in the new R01 for tumor tissue collection.

7. The P01 mechanism has some applicability in that it supports shared core resources as well as a few hypothesis testing projects. However, P01 applications face the same struggles for funding as R01s (including needing to focus mainly on the proposed science and the differing perspectives of a specific set of reviewers). In addition, because the core resources must be closely aligned with the specific scientific projects, some important infrastructure activities may be left out.

With regard to funding of new cohorts, the Working Group discussed whether applications for new cohorts should be evaluated in the same study section as applications for continuation of cohorts, or whether there should be a special cohort infrastructure study section for new cohorts. One possibility is to have an open call for new cohort applications on a periodic basis, rather than reviewing new unsolicited applications in an uncoordinated manner. The NIH also has a developmental funding mechanism that could be considered for new cohorts. Under this mechanism, new cohorts would have 2-3 years to demonstrate feasibility.

Concern was expressed as to whether the U01 funding pool is sufficient to fund new cohorts while maintaining existing ones. Discussion of new cohorts led to discussion of "sunsetting," i.e., when funding for a cohort should end or greatly diminish. Analyses from NCI indicate that cohort costs do not decrease over time after completion of initial enrollment, despite declines in follow-up due to deaths. Reasons include rising costs of technology, mailings, and sample collection, processing and storage; the inclusion of creative extras that are required to be attractive to study sections; and the greater number of cancers as the cohort ages, requiring more effort to obtain medical records and tumor tissue. The productivity in terms of research grants and publications varies by cohort, with the number of publications increasing over time from initiation of the cohort.

There was a consensus that age of participants should not be the only criteria used for changes in funding for a given cohort. The key consideration should be the cohort's potential to continue to contribute important scientific and public health information as assessed by the peer review process (see Question 1: Recommendations and Opportunities for Enhancement). The Working Group noted that PIs of the Iowa Women's Health Study chose to sunset it after a significant percentage of their participants had died, concluding that they could no longer justify continuing active follow-up of the cohort.

It was agreed, that in most instances, sunsetting may be more of a winding down process rather than an abrupt cessation of funding. Gradual steps include discontinuation of data collection with continuation of passive follow-up through cancer registries and the National Death Index (NDI), discontinuation of even passive follow-up, and creation of comprehensive fixed data files that

can be used in the future with little or no additional cost. The Working Group agreed that is also important that a structure remain in place for continuation of data analysis and specimen use.

Recommendations and Opportunities for Enhancement

1. The NCI should continue to use a Cohort Infrastructure Program Announcement for funding infrastructure activities of cancer cohorts. Investigator-initiated hypothesis-driven research based on cohort data would continue to be funded through R grants, P01s, and related mechanisms.
2. Applications for new cohorts should be considered in a special study section, separate from the study section that reviews continuations of cohorts.
3. It may be most effective for the NCI to accept applications for new cohorts only in response to a call for applications, which would occur periodically as needed.
4. Decisions about when to stop funding active follow-up of a given cohort should be made based on the likely productivity and importance of findings that will occur over the next five years.
5. A specific subheading could be added to the cohort infrastructure application to ensure that PIs will give a detailed rationale and justification for continuation of the cohort for another five years.
6. There is a need for further discussion to determine best practices for *whether* and *how* samples should be preserved for future use after funding for a given cohort has ceased, as well as who will make decisions about biospecimen use. At a minimum there is a cost to keeping freezers operating and supporting sample management.

Appendix I: Tables

Table 1. Extramural epidemiology cohorts[1] currently supported by NCI (as of 4/30/19)

Cohort	Cohort Abbrev	Year of Initial NCI Funding	Contact Principal Investigator[2]	Current Award (Annual Direct+Indirect Costs)[3]
Risk Cohorts				
Atherosclerosis Risk in Communities Cancer Cohort	ARIC-Ca	2012	Elizabeth Platz	$958,253
Black Women's Health Study	BWHS	1995	Lynn Rosenberg	$3,131,894
Breast Cancer Family Registries	Breast CFR	1996	Mary Beth Terry	$1,998,126
Breast Cancer Surveillance Consortium	BCSC	1994	Diana Miglioretti	$3,475,469
California Teachers Study	CTS	1998	James Lacy	$2,520,260
Carotene & Retinol Efficacy Trial[4]	CARET	1985	Chu Chen	$546,507
Colon Cancer Family Registries	Colon CFR	1998	Mark Jenkins	$2,128,492
Health Professionals Follow-up Study	HPFS	1986	Walter Willett	$2,002,569
Multiethnic Cohort	MEC	1986	Loic Le Marchand	$3,418,975
Nurses' Health Study	NHS	1976	Meir Stampfer	$2,867,003
Nurses' Health Study II	NHS II	1989	Walter Willett	$2,423,825
NYU-Women's Health Study	NYU-WHS	1985	Anne Zeleniuch-Jacquotte	$701,162
Prostate Cancer Prevention Trial/ Selenium & Vitamin E Cancer Prevention Trial[4]	PCPT/SELECT	1993 (PCPT) 2001 (SELECT)	Catherine Tangen	$908,559
Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial[5]	PLCO	1992	Paul Pinsky; Neal Freedman;	Intramural and Extramural Program funds
Shanghai Men's Health Study	SMHS	2001	Xiao Shu	$958,253
Shanghai Women's Health Study	SWHS	1996	Wei Zheng	$1,268,648
Shanghai Cohort Study/Singapore Chinese Health Study	SSC	1986 (Shanghai) 1993 (Singapore)	Jian-Min Yuan	$846,900
Southern Community Cohort Study	SCCS	2001	William Blot	$2,773,932
Women's Health Study[4]	WHS	1993	Julie Buring	$731,142
Survivor Cohorts				

Bone Marrow Transplant Survivor Study II	BMTSS-II	2019	Smita Bhatia	$1,312,128
Boston Lung Cancer Survival Cohort	BLCSC	2017	David Christiani	$1,626,509
Childhood Cancer Survivor Study	CCSS	1993	Gregory Armstrong	$4,120,774
ColoCare Study (Colorectal Cancer)	ColoCare	2016	Cornelia Ulrich	$1,745,632
Detroit Research on Cancer Survivors	Detroit ROCS	2017	Ann Schwartz	$1,952,390
Lymphoma Epidemiology of Outcomes Cohort	LEO	2015	James Cerhan	$2,246,690
WHI Life and Longevity after Cancer Study	LILAC	2013	Garnett Anderson	$2,260,718
Pathways (Breast Cancer)	Pathways	2004	Lawrence Kushi	$2,012,287
Research on Prostate Cancer in Men of African Ancestry	RESPOND	2018	Christopher Haiman	$3,226,742
St Jude LIFE Study (Childhood Cancer)	SJLIFE	2015	Melissa Hudson	$1,779,144

[1] - Defined as the cohort receiving an award through a cohort infrastructure grant mechanism that required thousands of study participants or a complex award mechanism that included a cohort infrastructure (i.e., RESPOND or BCSC cohorts)

[2] – Many of the grants are multi-Principal Investigator awards; only the Contact Principal Investigator is listed

[3] – Includes direct and indirect costs for the cohort infrastructure

[4] – Initially funded as a cancer prevention trial

[5] – Initially funded as a cancer screening trial

Table 2: Number of study participants by race and ethnic group – risk cohorts[1]

Cohort[2]	Recruiting? (Y/N)	White	Black or African American	Asian	American Indian/ Alaska Native	Native Hawaiian/ Pacific Islander	Total All Races	Hispanic[3]
U.S. Cohorts								
ARIC-Ca	No	11,478	4,266	34	12	--	15,790	230
BWHS	No	--	59,050	--	--	--	59,050	--
Breast CFR	Yes	31,520	2,323	2,319	99	44	36,305	3,803
BCSC	No	1,374,770	230,000	160,000	--	--	1,764,770	100,000
CTS	No	121,201	3,548	3,638	1,310	925	130,622	5,405
CARET	No	17,067	530	207	149	--	17,953	275
Colon CFR	No	31,969	1,946	2,188	329	115	36,547	484
HPFS	No	50,168	531	830	--	--	51,529	--
MEC	No	96,797	35,107	56,921	--	13,971	202,796	47,438
NHS	No	116,823	2,563	912	54	19	120,371	1,221
NHS II	No	111,671	2,305	2,438	--	--	116,414	1,878
NYU-WHS	No	9,866	1,489	115	--	--	11,470	771
PCPT/SELECT	No	43,045	5,962	550	185	56	49,798	2,793
PLCO	No	132,578	7,705	5,576	389	801	147,049	171
SCCS	No	25,378	55,578	107	329	--	81,392	230
WHS	No	37,822	917	546	103	--	39,388	430
TOTAL		**2,212,153**	**413,820**	**236,381**	**2,959**	**15,931**	**2,881,244**	**165,129**
TOTAL WITHOUT BCSC[4]		**837,383**	**183,820**	**76,381**	**2,959**	**15,931**	**1,116,474**	**65,129**
Asian Cohorts								
SMHS	No	--	--	61,491	--	--	61,491	--
SWHS	No	--	--	75,220	--	--	75,220	--
SSC	No	--	--	81,501	--	--	81,501	--
TOTAL		**--**	**--**	**218,212**	--	--	**218,212**	--

[1] – From the last competing grant application

[2] – Cohort acronyms explained in Table 1

[3] – Hispanic ethnicity ascertained separately from race

[4] – BCSC (funded through a P01), the largest cohort, is removed from this total because breast cancer is the only outcome and limited exposure date were obtained

Table 3: Number of study participants by race and ethnic group – survivor cohorts[1]

Cohort[2]	Recruiting? (Y/N)	White	Black or African American	Asian	American Indian/ Alaska Native	Native Hawaiian/ Pacific Islander	Total All Races	Hispanic[3]
BMTSS-II	Yes	8,123	451	181	45	45	8,845	993
BLCSC	Yes	9,387	312	434	10	795	10,938	--
CCSS	No	22,695	1,710	376	127	10	24,918	1,525
ColoCare	Yes	3,299	516	203	39	24	4,081	294
Detroit ROCS	Yes	278	8,062	--	--	--	8,340	--
LEO	Yes	3,920	58	38	21	--	4,037	97
LILAC	No	12,146	662	203	43	--	13,054	235
Pathways	No	3,177	300	554	15	6	4,052	565
RESPOND[4]	Yes	--	--	--	--	--		--
SJLIFE	Yes	6,635	1,558	44	5	3	8,245	--
TOTAL		**69,660**	**13,629**	**2,033**	**305**	**883**	**86,510**	**3,709**

[1] – Planned enrollment from last competing grant application

[2] – Cohort acronyms explained in Table 1

[3] – Hispanic ethnicity ascertained separately from race

[4] – Has not started recruiting

Table 4: Cancer Sites Reported in Risk and Survivor Cohorts – From the Cancer Epidemiology Descriptive Cohort Database (CEDCD; https://cedcd.nci.nih.gov/)[1]

	Male		Female	
Breast	300	0%	68,378	64%
Prostate	28,236	29%	N/A	N/A
Colon & rectum	15,645	16%	21,193	20%
Lung & bronchus	11,737	12%	15,450	14%
Non-Hodgkin lymphoma	8,043	8%	9,708	9%
Ovary	N/A	N/A	5,685	5%
Uterine corpus	N/A	N/A	9,255	9%
Urinary bladder	4,604	5%	2,510	2%
Melanoma	2,869	3%	6,503	6%
Leukemia	2,694	3%	3,776	4%
Liver & intrahepatic bile duct	2,492	3%	1,366	1%
Pancreas	2,406	2%	3,646	3%
Esophagus	1,333	1%	579	1%
Brain & other nervous system	1,023	1%	1,526	1%
Thyroid	522	1%	3,028	3%
All Other Sites	16,421	17%	23,297	22%
TOTAL CANCER COUNT	**98,025**		**107,522**	

[1] – The Cancer Epidemiology Descriptive Cohort Database includes centralized information about most of the cohorts as required by a funding mechanism that supports those cohorts. Data as of 2017.

Table 5: Number and type of biospecimens available from risk and survivor cohorts – From the Cancer Epidemiology Descriptive Cohort Database (CEDCD; https://cedcd.nci.nih.gov/)[1]

Cohort	Serum/ Plasma	Saliva/ Buccal	Urine	Feces	Buffy Coat	Tumor Tissue
Risk Cohorts						
ARIC-Ca	--	--	--	--	--	3,475
BWHS	13,000	26,800	--	--	13,000	--
Breast CFR	15,712	1,481	--	--	6,815	4,942
BCSC[2]	--	--	--	--	--	--
CARET	19,668	--	--	--	--	451
CTS	35,500		--	--	35,500	460
Colon CFR	29,636	491	--	--	27,211	10,738
HPFS	--	13,845	--	--	18,102	4,635
MEC	74,585	848	39,486	6,225	74,582	1,102
NHS	32,822	33,100	39,486	--	32,822	11,181
NHS II	29,611	29,850	46,121	--	29,611	--

NYU-WHS	15,026	--	--	--	--	449
PCPT/SELECT	53,666	--	--	--	43,404	3,452
SCCS	39,128	38,090	23,450	--	39,128	1,121
SMHS	49,688	8,467	54,349	--	--	1,546
SWHS	60,232	8,622	63,897	--	--	2,455
SSC	43,978	9,304	49,440	--	28,787	--
WHS	26,188	--	--	--	26,188	--
TOTAL	538,440	170,898	316,229	6,225	375,150	46,007
Survivor Cohorts						
BMTSS-II[3]	--	--	--	--	--	--
BLCSC[3]	--	--	--	--	--	--
CCSS[3]	--	--	--	--	--	--
ColoCare	1,432	--	355	384	--	1,809
Detroit ROCS	--	428	--	--	--	50
LEO	6,778	--	--	--	6,337	2,595
LILAC	--	--	--	--	--	4,105
Pathways	4,202	4316	--	--	4,202	--
RESPOND[4]	--	--	--	--	--	--
TOTAL	12,412	4,744	355	384	10,539	8,559

[1] – Cohort acronyms explained in Table 1. The Cancer Epidemiology Descriptive Cohort Database includes centralized information about most of the cohorts as required by a funding mechanism that supports those cohorts. Data as of 2017.

[2] – Did not collect biospecimens

[3] – Biospecimen data on cohorts not available in Cancer Epidemiology Descriptive Cohort Database

[4] – Has not started recruiting

Appendix II: List of Expert Presentations to the Working Group

The following talks were presented to the Working Group between September 2018 and January 2019.

- James McClain, Acting Chief Technology Officer, *All of Us* [SM] Research Program, NIH
  Presentation: "An Update on the All of Us Research Program"

- Montserrat Garcia-Closas, Deputy Director, Division of Cancer Epidemiology and Genetics (DCEG), NCI
  Presentation: "Prospective Cohort within Integrated Health Care Systems: Investment for the Future"

- Castine Clerkin, Virtual Pooled Registry Manager, North American Association of Central Cancer Registries (NAACCR)
  Presentation: "Virtual Pooled Registry Cancer Linkage System"

- Lynne Penberthy, Associate Director, Surveillance Research Program, Division of Cancer Control and Population Sciences (DCCPS), NCI
  Presentation: "The Evolution of SEER"

- Kathy Helzlsouer, Associate Director, Epidemiology and Genomics Research Program, DCCPS, NCI
  Presentation: "Cohort PAR Assessment Project"

Appendix III: Description of Major U.S. Cancer Epidemiology Cohorts NOT Supported by the NCI Extramural Program

Beyond the current NCI funded portfolio, there are several major intramural initiatives, including the *All of Us*[SM] Research Program and the NCI's *Connect* Cohort, that could potentially be a resource for the extramural community to utilize in examining the determinants of cancer and outcomes after a cancer diagnosis and need to be considered within the context of future needs relative to cancer cohorts.

All of Us[SM] is a major NIH initiative to establish a cohort of one million persons residing in the U.S. At this time, it has accrued over 100,000 people. People enrolled in the study respond to questionnaire surveys about health-related matters, receive a limited physical examination, and provide biospecimens and access to their EHRs. Cohort members are enrolled through a number of health provider organizations or directly volunteer through an enrollment portal. The cohort will be followed indefinitely through various means such as ongoing EHR acquisition, repeated

questionnaire surveys, additional biospecimen acquisition, mobile health technology data acquisition from study participants, linkage to the NDI and planned linkage to cancer registry data. The study has many substantial strengths, such as a high proportion of racial, ethnic, and underserved populations; innovative technologic approaches; planned return of individual results; and extensive outreach and engagement. The program plans to extensively share not only the data, but experiences, processes, instruments, tools, and software. For example, they are partnering with others to develop Sync for Science, a tool that will enable study participants to direct electronic health data directly to the *All of Us*SM Research Data Center and would share this experience with other researchers. It is not clear if the cohort is adequately powered for cancer survivor studies or to examine the risk of rare cancers.

*All of Us*SM plans to enable NIH Institutes to co-fund additional data collection from cohort members or enable *All of Us*SMcohort members to participate in special studies. Therefore, it is possible that NCI could support cancer-focused data collection from study participants or their health care providers to address cancer risk and outcome studies; e.g., for NCI to fund acquisition of tumor tissue from study participants and support linkage of the *All of Us*SM cohort to the Virtual Pooled Cancer Registry to ascertain incident cancer data.

Data will be available for analysis by any researcher, and cancer risk and outcome researchers will be able to use the data to research topics related to cancer risk and outcomes. However, there are some potential limitations of *All of Us*SM. The extent to which the *All of Us*SM Research Program will be able to retain study participants is not yet clear. The frequency of biospecimen collections has not been established, and there is likely to be substantial competition to use the biospecimens. The cohort is intended to be a research resource for the researcher stakeholders of all 27 NIH Institutes. Competition could be very stiff for access to study participants for special topics and respondent burden and burn-out could limit how much additional information *All of Us*SM can obtain from people beyond the data collection items that are broadly applicable to the health research community that the Program plans to collect from the one million participants.

The NCI's *Connect* Cohort is an initiative of the NCI Intramural Program to establish a new cancer epidemiology cohort; it is intended to be a major research resource for both the extramural and the NCI Intramural researcher communities. Approximately 200,000 persons will be ascertained from numerous large health care provider organizations across the U.S. The organizations were selected to collectively cover a diversity of regions across the U.S. and to include population diversity in aggregate, although there are no specific population subgroup enrollment targets. The investigators will obtain participant-provided information about behaviors, family history, and many other characteristics; collect electronic health records continuously; and obtain a variety of biospecimens from study participants, including tumor tissue. The investigators expect to re-survey and collect biospecimens on a regular basis over the life of the cohort. Follow-up methods and endpoints include linkage to cancer registries via the Virtual Pooled Cancer Registry and linkage to the NDI for incidence, survival, and mortality. They also plan to collect data about recurrences. The data will be made available to the

extramural community. Some innovations that will benefit the entire research community are the plans for storing all the metadata about the protocol and data, accessing, and analyzing the data in the cloud. It is not clear if the cohort is adequately powered for survivorship studies, risk of rare cancers, or less common racial and ethnic groups.

There are also large, non-NCI-funded cancer epidemiology risk and survivor cohorts supported by other entities. The American Cancer Society (ACS) recruited a new prospective cohort study, Cancer Prevention Study 3 (CPS-3), between 2006 and 2013 from 35 states and Puerto Rico (Patel, 2017). Enrollment took place primarily at ACS community events and at community enrollment "drives." At enrollment sites, participants completed a brief survey that included an informed consent, identifying information necessary for follow-up, and key exposure information. They also provided a waist measure and a nonfasting blood sample. Most participants also completed a more comprehensive baseline survey at home that included extensive medical, lifestyle, and other information. Participants will be followed for incident cancers through linkage with state cancer registries and for cause-specific mortality through linkage with the NDI. In total, 303,682 participants were enrolled. Of these, 254,650 completed the baseline survey and are considered "fully" enrolled; they will be sent repeat surveys periodically for at least the next 20 years to update exposure information. The remaining participants (n = 49,032) will not be asked to update exposure information but will be followed for outcomes. Twenty-three percent of participants were men, 17.3% reported a race or ethnicity other than White, and the median age at enrollment was 47 years. Having completed the initial recruitment of participants, ACS investigators are shifting to the follow-up phase of the study, which includes the first follow-up survey for the entire population. In addition to the cohort-wide follow-up survey, a small group of CPS-3 participants are being invited to participate in one of two sub-studies on diet, physical activity, light, and sleep patterns. CPS-3 will be an important resource for studies of cancer and other outcomes because of its size; its diversity with respect to age, race/ethnicity, and geography; and the availability of blood samples and detailed questionnaire information collected over time.

These transformations of prevention and treatment trials generally were achieved through collection of cancer incidence data, survival, mortality, or physical health outcome ascertainment beyond the completion of the primary aims of the trial. The following types of data collection approaches in addition to those required by the trial were typically employed: questionnaires to study participants about occurrence of cancer and outcomes, typically followed by medical record abstraction, linkage with cancer registries and Medicare files, and NDI linkages. Many of these prevention and treatment trials have special features that make them particularly valuable, including very large sample size (for the prevention studies), repeated biospecimen collections on all study participants, tumor tissue, and collection of other data at regular time intervals (e.g., dietary intake and tobacco and alcohol use).

References

Bluethmann SM, Mariotto AB, Rowland, JH. Anticipating the "Silver Tsunami": Prevalence Trajectories and Comorbidity Burden among Older Cancer Survivors in the United States. *Cancer Epidemiol Biomarkers Prev*. 2016;25:1029-1036.

Cronin KA, Lake AJ, Scott S, et al. Annual Report to the Nation on the Status of Cancer, part I: National cancer statistics. *Cancer*. 2018;124(13):2785–2800.

Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res*. 2014 Jun 1;74(11):2913-21.

U.S. Food and Drug Administration. Real World Evidence. https://www.fda.gov/ScienceResearch/SpecialTopics/RealWorldEvidence/default.htm, Accessed May 9, 2019.

Patel AV, Jacobs EJ, Dudas DM, Briggs PJ, Lichtman CJ, Bain EB, Stevens VL, McCullough ML, Teras LR, Campbell PT, Gaudet MM, Kirkland EG, Rittase MH, Joiner N, Diver WR, Hildebrand JS, Yaw NC, Gapstur SM. The American Cancer Society's Cancer Prevention Study 3 (CPS-3): Recruitment, study design, and baseline characteristics. *Cancer*. 2017 Jun 1;123(11):2014-2024.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016 Mar 15;3:160018.

Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2019. *CA Cancer J Clin*. 2019;69:7-34.